ELSEVIER

Robotics and Autonomous Systems 12 (1994) 121–131

Robotics and Autonomous Systems

Attentive behavior in an anthropomorphic robot vision system

C. Colombo *, Michelle Rucci, Paolo Dario

Advanced Robotics Technology and Systems Laboratory (ARTS Lab), Scuola Superiore S. Anna, Pisa, Italy

Abstract

Powerful data reduction and selection processes, such as selective attention mechanisms and space-variant sensing in humans, can provide great advantages for developing effective real-time robot vision systems. The use of such processes should be closely coupled with motor capabilities, in order to actively interact with the environment. In this paper, an anthropomorphic vision system architecture integrating retina-like sensing, hierarchical structures and selective attention mechanisms is proposed. Direction of gaze is shifted based on both the sensory and semantic characteristics of the visual input, so that a task-dependent attentive behavior is produced. The sensory features currently included in the system are related to optical flow invariants, thus providing the system with motion detection capabilities. A neural network architecture for visual recognition is also included, which produces semantic-driven gaze shifts.

Key words: Robot vision; Retina-like sensing; Selective attention; Object recognition; Optical flow; Neural networks

1. Introduction

An enormous amount of information falls at every instant onto the eyes; nevertheless, vision in humans appears to be effortless. At least two major factors account for this effectiveness: the *space-variant* structure of human retina [1], and the mechanism of *selective attention,* that is the capability of selectively processing simultaneous sources of visual information [2]. The two factors cooperate to dramatically reduce the computational burden at the retina and the visual cortex levels, respectively. Similar complexity-reduction mechanisms should be taken into consideration when attempting to develop an effective real-time robot vision system.

Selective attention in humans has been extensively studied in the past decades. Experiments in psychophysics have proved that selective attention in visual perception has the characteristic of a limited extent attentional spotlight [3]. The spotlight can be voluntarily moved independently on eye fixations, can vary in size, and is drawn by significant stimuli in the scene, thus taking part in the control of rapid shifts of gaze direction, or *foveations.* Shifts of attention and eye movements [4] seem to play an important role for visual exploration and recognition [5,6]. A major issue of research on human vision is the understanding of which factors contribute to shift attention on specific parts of the visual field. A fundamental

* Corresponding author.

0921-8890/94/$07.00 © 1994 Elsevier Science B.V. All rights reserved
SSDI 0921-8890(93)E0037-N

fact, already pointed out in 1890 by William James [7], is that visual attention can be drawn both by the *sensory* and *semantic* characteristics of a stimulus, i.e. both by bottom-up and top-down visual mechanisms.

In the field of robot vision, several system architectures have been proposed which use space-variant sensing strategies in order to reduce the raw data being processed. These strategies can be grouped into two categories, namely *artificial retinae* [8,9] and *hierarchical structures* [10,11]. The space-variant sensing structure of artificial retinae is a direct consequence of the retina-like geometry of the sensor, which is composed of a high-resolution, small central *fovea* and a *periphery* whose resolution linearly decreases with eccentricity. As for the human visual system, only one foveation at a time can be carried out with artificial retinae.

A hierarchical structure or *pyramid* is a representation of a rectangular image – e.g. the output of a traditional high-resolution camera – at increasingly lower resolution levels. Foveations with hierarchical structures are the result of a coarse-to-fine search through certain paths of the pyramid; space-variant sensing in pyramids is thus not accomplished at a 'hardware' but at an 'algorithmic' level, thus allowing in principle more than one (simulated) foveation to occur in parallel.

With the noticeable exception of the work done on pyramidal structures (e.g. [12,13]) by some research groups, few research experiences exist so far in robot vision on the topic of selective attention. Moreover, research was mainly based on traditional high-resolution cameras [14], thus not fully exploiting the possibilities of space-variant sensing.

In this paper, we propose an anthropomorphic architecture for a robot vision system. The architecture integrates retina-like sensing, hierarchical image analysis and selective attention mechanisms in order to produce an attentive visual behavior. Active control of gaze is achieved in the system by suitably combining both the semantic and sensory characteristics of the visual scene on the basis of an attended task. That is, the way the system interacts with the environment is task-dependent; from this point of view, the architecture proposed here can be included in the *active vision* paradigm [15].

In the present implementation, special interest has been devoted to motion features at the sensory level and to object recognition at the semantic level, so that the system is able to foveate and

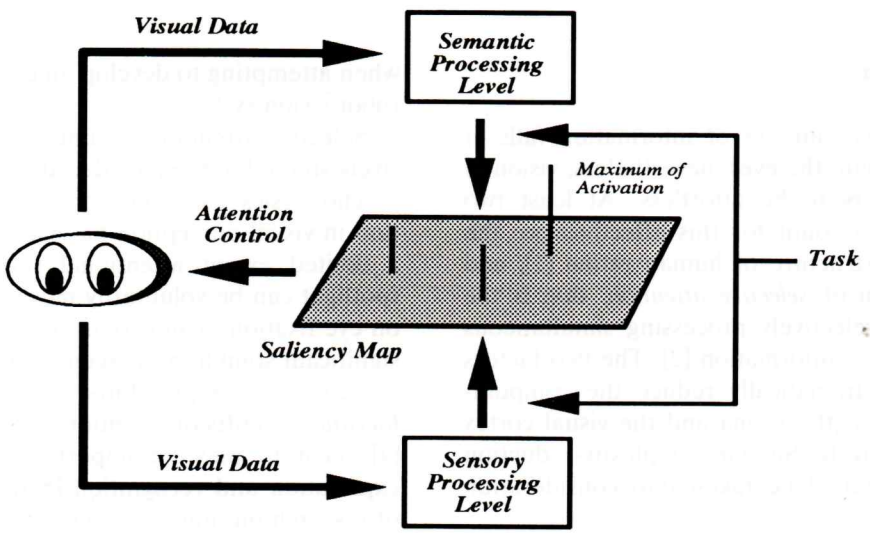


Fig. 1. The system architecture. The sensory and semantic processing levels operate in parallel on the data produced by the sensor. Next foveation occurs where the saliency map exhibits a maximum of activation.

explore moving objects, and to recognize them if they are already known. Thanks to its intrinsic modularity, the architecture is suitable to the expanded so as to include different tasks and also other sensory modalities, such as touch or hearing.

The paper is organized as follows: in Section 2 the architecture is presented and described, then in Section 3 experimental results are discussed, and finally in Section 4 conclusions are drawn.

2. Overview of the system

A general scheme of the system architecture is shown in Fig. 1. The system includes an anthropomorphic sensor, which is given two d.o.f. and moves so as to explore the environment as the result of an attentive behavior.

The logical center of the architecture is the attention control mechanism; based both on the selected task and on the results of an analysis of the characteristics of the input, attention control delivers the proper motion commands to the actuation subsystem so as to shift gaze to the desired visual location. This is accomplished by letting each sensory (bottom-up) process run in parallel with the others and contribute to activate corresponding specific locations of a *saliency map* [16]. The map can also be activated in a top-down fashion by the selection of specific semantic cues stored in memory. The mapping onto the saliency map is performed by a set of priority-ordered connections gated by the task at hand – that is, the relative weight assigned to each map input is task-dependent. Notice that no distinction is made between semantic and sensory map inputs, so that the total saliency map input is simply the superposition of the weighted inputs from the bottom-up and top-down processes. Each new foveation takes place in the direction where saliency map activation reaches its maximum; as a result, attentional shifts are always accompanied in this system by corresponding movements of the sensor or, in other words, the 'focus of attention' is always kept centered on the fovea.

A detailed description of the single elements of the architecture follows.

2.1. Attention control and the saliency map

Attentional shifts are controlled by communicating to the actuation subsystem the saliency map location which currently exhibits the maxi-

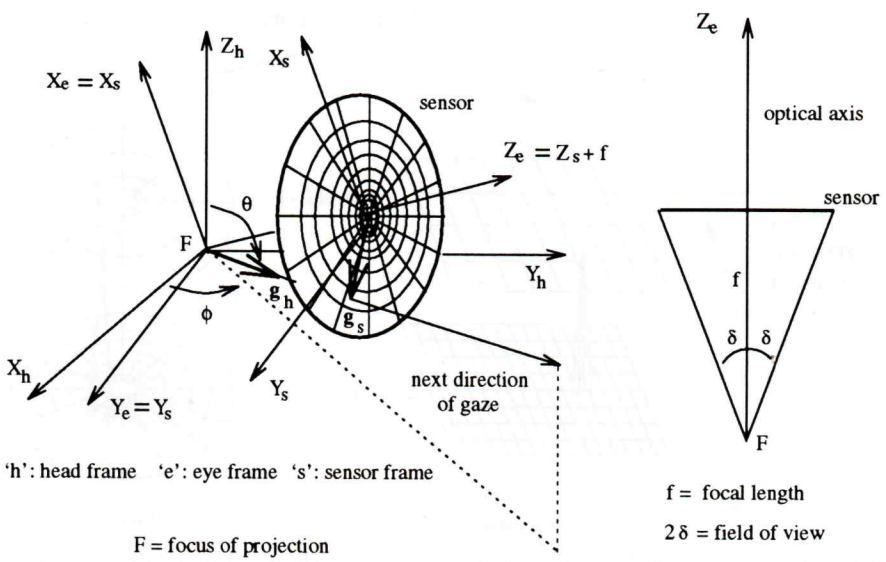


Fig. 2. Visual directions and the geometry of the head–eye system. For the sake of simplicity, the head (fixed) and the eye (mobile) frame origins have been assumed coincident. The field of view is equal to 2 arctan (a^{M+p}/f) (see text).

mum activation. This location can be varied by assigning a different priority to the corresponding sensory/semantic characteristic, thus modifying the current task.

A saliency map location is simply a visual direction, parameterized by the angles θ (colatitude) and ϕ (longitude) relative to a head frame centered on the focus of perspective projection (see Fig. 2). It can be alternatively represented as the unit homogeneous vector

$$\boldsymbol{g}_h = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta, 1). \quad (1)$$

The mapping of a visual direction from its representation $\boldsymbol{g}_s$ in the sensor frame – that provided by both the input processing levels - to the corresponding saliency map representation $\boldsymbol{g}_h$ can be written as

$$\boldsymbol{g}_h = {}^{e}_{h}\boldsymbol{H}\,{}^{s}_{e}\boldsymbol{H}\,\boldsymbol{g}_s, \quad (2)$$

where the matrices ${}^{e}_{h}\boldsymbol{H}$ and ${}^{s}_{e}\boldsymbol{H}$ are homogeneous transformations between the eye and head frames, and the sensor and eye frames, respectively. These matrices have entries which depend on the focal length f and on 'proprioceptive' data obtained from the position sensors of the actuation subsystem.

The saliency map is non-retinotopic; it follows that the set of gaze directions explored by the visual sensor at a given time is a subset of the map, whose size depends on the field of view. In principle, also the use of other sensory modalities (touch, hearing) and/or semantic cues (abstract representations of the environment and of the task) can determine a gaze shift to visual directions which are currently out of sight.

2.2. The sensory processing level

Concerning the sensory characteristics, space-variant sensing is performed both at the sensor and at the algorithmic level by means of a hierarchical space-variant representation of the scene.

The structure of a retina-like sensor and the pyramid construction process from a generic level L are depicted in Figs. 2 and 3, respectively. At the lowest level of the hierarchy – $L = 0$, or sensor level – the periphery around the central fovea is partitioned in M annuli $\times N$ angular sectors. The ratio of the outer and inner radii of each annulus is equal to a constant $a > 1$.

In order to build the pyramid, a rectangular representation of the space-variant sensed image

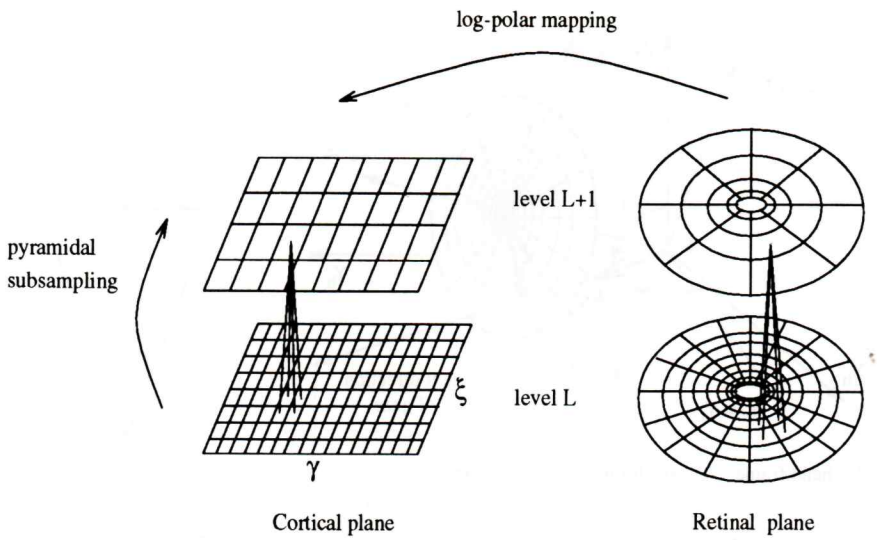


Fig. 3. Left: building binary cortical pyramids. As one more level in the hierarchy is added, cortical images are first smoothened by a 5×5 Gaussian filter and then subsampled by a factor of two. Right: backtransforming a cortical pyramid yields a retinal pyramid: notice the presence of a hollow inner region corresponding to the fovea.

is obtained by letting the generical periphery location at polar coordinates (ρ, ϕ), undergo the 'log-polar' transformation [17]

$$\begin{cases} \xi = \log_a \rho - p \\ \gamma = q\eta, \end{cases} \tag{3}$$

where $p = \log_a \rho_{\text{fovea}}$ and $q = N/2\pi$. Binary pyramids are built in the *cortical plane* (ξ, η) according to a well-known algorithm [18]; i.e., pyramid level L, $L \geqslant 1$, is built by first filtering and then subsampling by a factor of two in both visual directions the cortical image at level $L-1$. Notice from Fig. 3 that each level $L+1$ of the *cortical pyramid* can be interpreted in the retinal plane as a novel – and coarsely sampled – retina, with characteristic parameters, $M_{L+1} = M_L/2$, $N_{L+1} = N_L/2$, $a_{L+1} = a_L^2$, $p_{L+1} = p_L/2$, $q_{L+1} = q_L/2$.

Sensory pyramids are obtained from a further processing of this cortical image brightness pyramid. Data in sensory pyramids are explored in a top-down fashion, i.e., Winner-Takes-All processes operate at each level of the pyramid so that only one 'winner' is propagated, while 'loosers' are inhibited [13]. In such a way computations are greatly reduced, as only small parts of the visual data are processed at higher resolutions, the rest being explored only at the lowest resolution.

The problem of computing motion sensory features – specifically, the invariants of the optical flow [19] – at the generical cortical pyramid level is now discussed. (The $_L$ is omitted from the retina parameters for simplicity.)

Although all relative to the retinal plane, motion features are computable by means of a transformed optical flow field computed in the cortical plane, which is referred to as *cortical flow*. The cortical flow is defined as the solution $(\dot{\xi}, \dot{\gamma})$ of the equation

$$\frac{\partial E}{\partial \xi}\dot{\xi} + \frac{\partial E}{\partial \gamma}\dot{\gamma} + \frac{\partial E}{\partial t} = 0, \tag{4}$$

where $E(\xi, \gamma, t)$ is the cortical image brightness, and can be obtained by a least squares technique over a neighborhood of each cortical pixel, under the assumption of a uniform speed in it [20].

Optical flow invariants – i.e. *magnitude* and the 'differential invariants' *divergence, curl,* and *shear* – are scalar quantities which locally characterize a linear optical flow field, and remain unchanged after a rotation of the image coordinate system [21]. Specifically, optical flow magnitude gives the local intensity of an object displacement, while divergence indicates a change in area, curl accounts for rotational flow, and shear indicates a deformation (an expansion plus a contraction) along perpendicular directions of the retinal plane so that the area is unchanged.

Differential invariants are functions of the spatial derivatives of the optical flow – see for example [22] – and are related in a simple way to the 3D spatio-temporal evolution of the imaged scene as well as to viewer motion [23]. As an example, the divergence of the optical flow can be put in correspondence with a motion towards (expansion) or away from (contraction) the retinal plane, while curl can be related to eye cyclotorsions about a focal axis.

Using the log-polar transformation 3 it can be easily shown that optical flow invariants can be expressed in terms of cortical flow and its spatial derivatives as

$$|\boldsymbol{v}| = a^{\xi+p}\left[\left(\ln a\dot{\xi}\right)^2 + \left(\frac{1}{q}\dot{\gamma}\right)^2\right]^{1/2}, \tag{5}$$

and

$$\begin{cases} \operatorname{div}(\boldsymbol{v}) = 2\ln a\dot{\xi} + \dfrac{\partial\dot{\xi}}{\partial\xi} + \dfrac{\partial\dot{\gamma}}{\partial\gamma} \\ \operatorname{curl}(\boldsymbol{v}) = \dfrac{2}{q}\dot{\gamma} - q\ln a\dfrac{\partial\dot{\xi}}{\partial\gamma} + \dfrac{1}{q\ln a}\dfrac{\partial\dot{\gamma}}{\partial\xi} \\ \operatorname{shear}(\boldsymbol{v}) = \left[\left(\dfrac{\partial\dot{\xi}}{\partial\xi} - \dfrac{\partial\dot{\gamma}}{\partial\gamma}\right)^2 + \left(q\ln a\dfrac{\partial\dot{\xi}}{\partial\gamma} + \dfrac{1}{q\ln a}\dfrac{\partial\dot{\gamma}}{\partial\xi}\right)^2\right]^{1/2}, \end{cases} \tag{6}$$

where $\boldsymbol{v} = \dfrac{\mathrm{d}}{\mathrm{d}t}(\rho\cos\phi, \rho\sin\phi)$ is the optical flow vector.

Another useful motion-related sensory feature is *time to impact* τ, that is the time for an object imaged at a certain retinal location to reach the observer. For a sufficiently small field of view, the time to impact is shown to be independent on eye/object rotations, and bounded by linear combinations of divergence and shear [24]:

$$\tau \approx \frac{2}{\mathrm{div}(\boldsymbol{v}) \pm \mathrm{shear}(\boldsymbol{v})}. \tag{7}$$

The motion sensory features described above can be effectively used as attentional features for robots. In Section 3, some experiments are described involving the optical flow magnitude and the time to impact sensory pyramids. Although a bit noisy, such sensory features are robust enough for drawing attention, which is mostly a *qualitative* problem, in that the relative intensity of a feature at different locations is more important than its absolute value.

2.3. The semantic processing level

The semantic processing level of the architecture contributes to controlling attention, during recognition, by providing salient spatial locations of the visual field which are relevant from a 'cognitive' point of view. Stored knowledge of the world is crucial for orienting attention in many cases. As psychophysical studies on humans have elucidated, attention plays a major role in operations such as visual exploration and recognition, where hypotheses formulated on the basis of the incoming visual data produce expectations on where salient cues are supposed to be found.

The semantic processing level incorporates a fragmentary representation of the objects to recognize including spatial relationships among the object parts. An example of object representation is shown in Fig. 4: several different *feature units,* each representing a part of the object, are linked to a single *object unit* which represents the object itself. The spatial relationships specify how the parts are located one with respect to others, and their relative dimensions. Each feature unit acts as a cumulator, by storing and cumulating the activation provided by a *matching network,* which

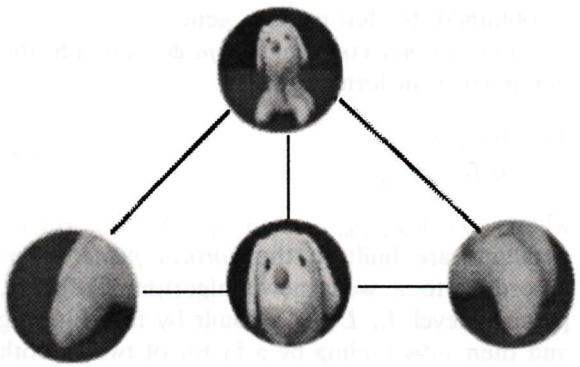


Fig. 4. The internal ionic representation of objects in the long-term memory. A feature unit sensitive to each icon is present in the system. Solid lines indicate the geometrical relationships among the various parts of the representation.

analyzes the data gathered from the scene through a limited extent attentional spotlight. All the object units inhibit each other in a Winner-Takes-All fashion, so that only one of them has a positive value of activation at a given time, while all the others are inhibited.

The spatial relationships are given in the retinal space, and they are normalized with respect to the object size, that is the dimension of the attentional spotlight when a feature unit is activated. In this way, the actual gaze shift to be performed to foveate to any other part of the object and the required spotlight size are evaluated by multiplying the stored normalized displacements and size by the current spotlight dimension. In the current implementation of the system, all the spatial relationships are specified with respect to the feature unit which is activated when the spotlight of attention includes the whole object. Thus, starting from each feature the parameters for the examination of each other part of the object are determined by passing through this unit.

The spatial parameters are estimated by a four-layer feed-forward net trained with the backpropagation algorithm [25]. The input to the network are the feature unit number and the position of the winning unit in a self-organizing map included in the matching network (see be-

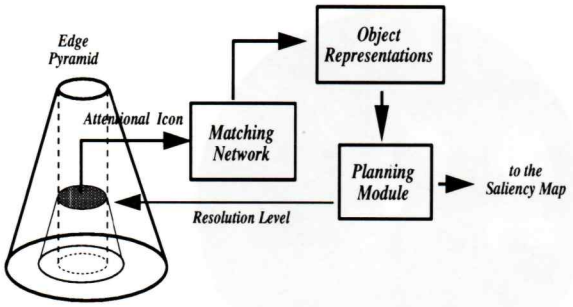


Fig. 5. The edge pyramid and the recognition task. This architecture takes inspiration on Nakayama's speculative framework of the overall structure of the human visual system [6].

low), and it produces as output, with a sparse coding, the coordinates of the spatial location of the feature and the spotlight dimension required for its examination.

The architecture of the semantic processing level is illustrated in Fig. 5 [26]. Image data are organized into a multi-resolution edge pyramid, which represents image edges at different levels of resolution. The pyramid is built on both data coming from the fovea and the periphery of the sensor. The edges are extracted at the lowest pyramid level by a gradient operator and data are then propagated at successive stages by means of Gaussian filtering. As shown in Fig. 5, the pyramid is scanned by an attentional spotlight centered on the fovea which, by moving through a fovea-centered cylinder, samples a fixed amount of information at different levels of resolution. As a result, a trade-off is built between the resolution level and the spatial extension of the considered area, and an increment of the width of the area implies a corresponding decrement of the level of resolution at which data are examined. The spotlight performs an expansion of the gray-level dynamics and produces a fixed dimension *attentional icon.* By means of the expansion, parts of the examined area with stronger edges and/or higher edge concentration are emphasized with respect to the others. When it is not involved in a recognition sequence, the spotlight scans the cylinder from the bottom to the top of the pyramid. This is equivalent to perform a radial expansion of the considered region starting from the fovea towards the periphery.

At any time, the resulting attentional icon is accepted as input by the matching net, which produces activation for the corresponding feature units. The matching net is a counterpropagation neural network [27] with as many output units as the feature unit number. Each output is linked with the corresponding feature unit in the object representation. The topological self-organizing map ([28]) at the second layer of the net stores all the icons used for the object representations.

The system performs recognition by serially looking at different parts and features of the examined object on the basis of the decision taken by a planning module. Recognition is achieved if the global input to the object unit is larger than a predetermined threshold. If a 'semantic match' is found [1] – that is, a hypothesis on the object identity is formulated by means of the activation of an object unit – the pyramid scanning process stops and the spotlight dimension is stored into a short-term memory so as to be used later for evaluating the spatial parameters of subsequent fixations.

In order to accept the hypothesis, the system begins a serial examination of the object features looking for other matches. If the hypothesis fails to be confirmed by successive attentional fixations, the planning module resets all the feature units in the representation of the rejected object. In this way, the second most probable hypothesis wins the competition and its features are then analyzed. The cycle is repeated until recognition is achieved or all the possibilities are sequentially examined.

3. Results

In the implementation described in this paper the system includes two pyramids at the sensory

[1] A 'Weak perspective' projection model [29] is assumed here, thus avoiding significant deformations of the peripheral parts of the foveated objects; such deformations could seriously complicate the ionic representation.

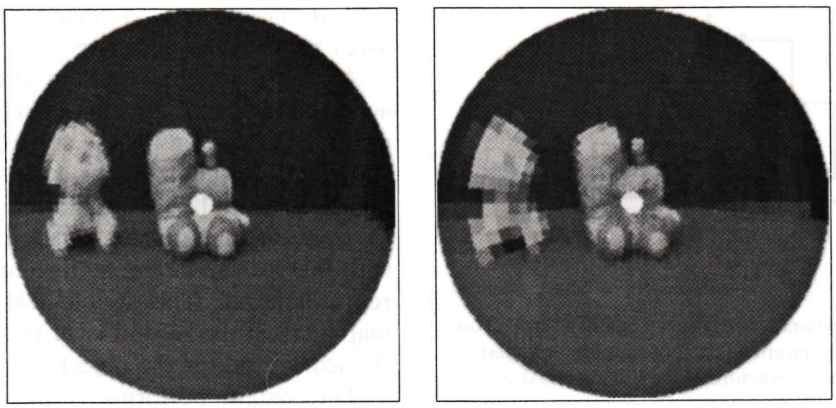


Fig. 6. Left: A frame as sensed by a retina-like sensor with $M = 64$ annuli and $N = 128$ angular sectors. The radii-ratio is $a = 1.04621$, and $p = 43.32067$. The field of view is approximately 45 deg. Only the periphery is shown. Right: The same frame, after cortical pyramidalization (level 1). Notice that image degradation is increasingly higher moving outwards from the fovea.

processing level and the representations of two objects in the semantic stage. The two pyramids represent the magnitude of the optical flow and the time to impact so that at any time the two locations with the highest values of these parameters are activated in the saliency map and can be selected by the system as possible directions of gaze.

The two objects known by the system are the dog and the train toys shown in Fig. 6. Both the objects have been represented with four feature units and a single object unit. The representations have been built by manually selecting the parts chosen in a number of images reproducing the objects with different orientations. The resulting icons were used for training the matching network and they were stored into the self-organizing map. The spatial relationships among the parts were used as the training set for the backpropagation-trained network. The output of an artificial retina has been simulated by resampling high-resolution 256×256 b/w images, upon which one pyramid level has been built.

An example showing the system at work is

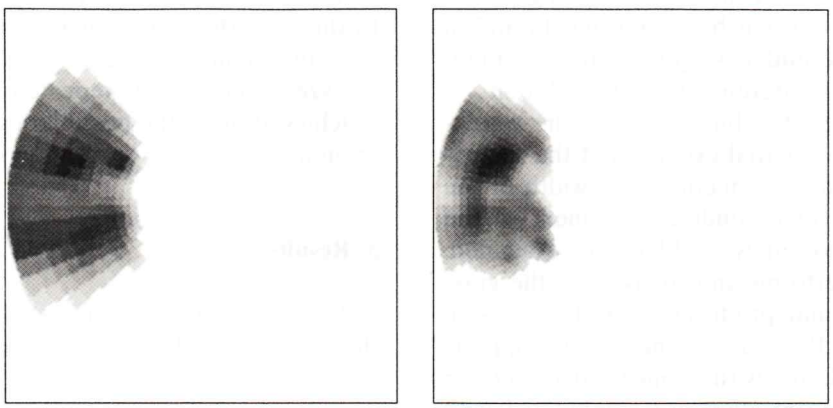


Fig. 7. Left: The lower bound on the time to impact as obtained from the differential invariants computed at the 1st level of the pyramid (darker areas correspond to closer surfaces). A 3×3 neighborhood was used to compute the cortical flow, and a 3×3 Gaussian filter was used to smoothen it. Right: The time to impact as it would look if the whole 0th level of the pyramid was processed. In practice, cortical flow – and consequently time to impact – is computed only at a few points at this level, specifically those correnspondig to the 1st level cortical pixel with lowest time to impact.

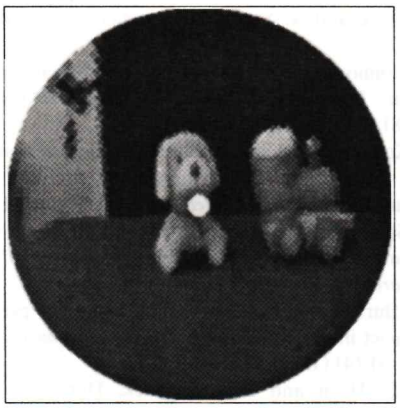


Fig. 8. The scene (sensor level) after foveation. Gaze has shifted on the muzzle of the dog, so the train is imaged now in the periphery of the retina.

illustrated in Figs. 6 through 8. The system is tuned so as to maximally attend to approaching motion – that is, time to impact pyramid has the maximum priority – followed by object recognition and by speed magnitude. Fig. 6 is the first frame of a 'stop-and-go' sequence involving the two toy-objects. At the beginning the system is looking at the train and it is involved in its recognition sequence. Then, the dog, which is imaged in the periphery of the visual field, takes a leap in the direction of the sensor. Finally, the train starts moving. As the dog moves, its approaching is detected by the time to impact pyramid, thus forcing the system to foveate in its direction. Due to the fact that a higher priority was assigned to the possibility of collisions with moving objects than to recognizing a still object, a nonzero entry is produced in the saliency map, and attention is drawn on the dog. Fig. 7 shows how a dog motion towards the retinal plane is perceived by the time to impact pyramid (notice that the train, which is still, is totally ignored by this pyramid). A first foveation (Fig. 8) occurs on the muzzle of the dog; it being the most prominent part of its body, it has produced the lowest time to impact. As the system 'knows' the dog, the recognition process starts, and a sequence of foveations takes place on the dog, so as to examine its different parts. Although the subsequent 'departure' of the train produces a cue in the magnitude speed pyramid (but not in the time to impact pyramid), the system is not 'distracted' from its recognition sequence, due to the higher priority of recognition.

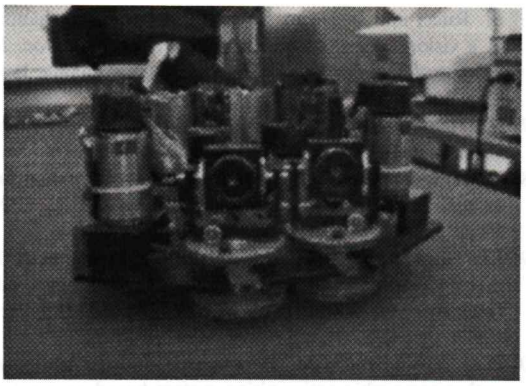


Fig. 9. The robotic head developed at the ARTS Lab. The system features a mechanical unit which includes two retina-like sensors, each actuated with two d.o.f. through DC servomotors, and a transputer-based control architecture [30].

4. Conclusions

Mechanisms for data reduction and selection are crucial for the development of effective robotic vision systems. Thanks to an attentive behavior, the computational resources of a system can be focused on the process of the relevant data, thus contributing to significantly improve the system performance. Experiments similar to the one described in the previous section show that the architecture could be effective as a part of a real-time vision system interacting with real-world environments. The integration of sensory and semantic characteristics of the visual data have produced an adaptive behavior which can change on the basis of the task at hand.

At present, the system architecture is being implemented on a binocular vision system recently developed at the ARTS Lab (Fig. 9). Due to the availability of two sensors, binocular sensory cues such as disparity are also being considered, together with light intensity, density of contours, and other features provided by different sensory modalities, such as hearing and touch. Also the learning capabilities of the system can be improved by means of on-line learning to-

wards the autonomous development of new object representations.

Acknowledgements

The work described in this paper has been supported, in part, by the Special Project on Robotics of the National Research Council of Italy. The authors wish to thank Dr. D.M. De Micheli for the image of the ARTS head. One of the authors (M. Rucci) has been supported by a fellowship from Istituto per la Ricostruzione Industriale.

References

[1] E.L. Schwartz, Spatial mapping in the primate sensory projection: Analytic structure and relevance to perception, *Biological Cybernetics* 25 (1977) 181–194.

[2] W.A. Johnston and V.J. Dark, Selective attention, *Annual Review of Psychology* 37 (1986) 43–75.

[3] M. Posner, Orienting of attention, *Quarterly Journal of Experimental Psychology* 32 (1980) 3–25.

[4] A.L. Yarbus, *Eye Movements and Vision* (Plenum Press, 1967).

[5] D. Noton and L. Stark, Eye movements and visual perception, *Scientific American* 224 (6), (1971) 34–43.

[6] K. Nakayama, The iconic bottleneck and the tenuous link between early visual processing and perception, in: C. Blakemore, ed., *Vision: Coding and Efficiency* (University Press, 1991).

[7] W. James, *The Principles of Psychology* (Harvard University Press, Cambridge, 1890/1983).

[8] C.F.R. Weiman and G. Chaikin, Logarithmic spiral grids for image processing and display, *Computer Graphics and Image Processing* 11 (1979) 197–226.

[9] G. Sandini and V. Tagliasco, An anthropomorphic retina-like structure for scene analysis, *Computer Graphics and Image Processing* 14 (3) (1980) 365–372.

[10] A. Rosenfeld, ed., *Multiresolution Image Processing and Analysis* (Springer, 1984).

[11] J.L. Crowley, A representation for visual information (Technical Report CMU-RI-TR-82-7, Carnegie-Mellon University, 1987).

[12] P.J. Burt, Smart sensing within a pyramid vision machine, *Proc. IEEE* 76 (8) (1988) 1006–1015.

[13] S.M. Culhane and J.K. Tsotsos, An attentional prototype for early vision, *Proc. 2nd European Conf. on Computer Vision,* S. Margherita Ligure, Italy (1992) 551–560.

[14] J.J. Clark and N.J. Ferrier, Attentive visual servoing, in: A. Blake and A. Yuille, eds., *Active Vision* (MIT Press, 1992).

[15] J. Aloimonos, I. Weiss and A. Bandyopadhyay, Active vision, *International Journal of Computer Vision* 1 (4) (1988) 333–356.

[16] C. Koch and S. Ullman, Shifts in selective visual attention: toward the underlying neural circuitry, In: L.M. Vaina, ed., *Matters of Intelligence* (D. Reidel, 1987).

[17] G. Sandini and P. Dario, Active vision based on space-variant sensing, *Proc. 5th Int. Symposium of Robotics Research,* Tokio (1989) 408–417.

[18] P.J. Burt and E.H. Adelson, The Laplacian pyramid as a compact image code, *IEEE Transactions on Communications* 31 (4) (1983) 532–540.

[19] B.K.P. Horn and B.G. Schunck, Determining optical flow, *Artificial Intelligence* 17 (1981) 185–203.

[20] F. Bartolini, V. Cappellini, C. Colombo and A. Mecocci, Multiwindow least squares approach to the estimation of optical flow with discontinuities, *Optical Engineering* 32 (4) (1993) 1250–1256.

[21] K. Kanatani, *Group-theoretical Methods in Image Understanding* (Springer, 1990).

[22] R. Cipolla and A. Blake, Surface orientation and time to contact from image divergence and deformation, *Proc. 2nd European Conference on Computer Vision,* S. Margherita Ligure, Italy (1992) 187–202.

[23] J.J. Koenderink and A.J. van Doorn, Invariant properties of the motion parallax field due to the movement of rigid bodies relative to an observer, *Optica Acta* 22 (9) (1975) 773–791.

[24] M. Subbarao, Bounds on time-to-collision and rotational component from first-order derivatives of image flow, *Computer Vision, Graphics, and Image Processing* 50 (1990) 329–341.

[25] D.E. Rumelhart, G.E. Hinton and R.J. Williams, Learning internal representations by error propagation, in: J.L. McClelland, D.E. Rumelhart and the PDP Research Group, eds., *Parallel Distributed Processing* (MIT Press, 1986).

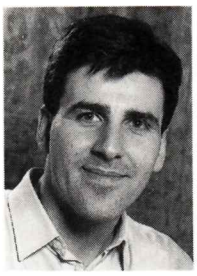
[26] M. Rucci and P. Dario, Selective attention mechanisms in a vision system based on neural networks, *Proc. Int. Conf. on Intelligent Robots and Systems*, Yokohama, Japan (July 1993).

[27] R. Hect-Nielsen, Applications of the counter-propagation networks, *Neural Networks* 2 (1) (1988).

[28] T. Kohonen, Self-organized formation of topologically correct feature maps, *Biological Cybernetics* 43 (1982) 59–69.

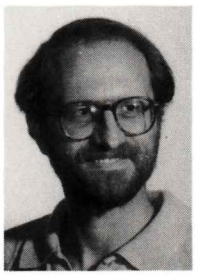
[29] J.L. Mundy and A. Zisserman, Projective geometry for machine vision, in: J.L. Mundy and A. Zisserman, eds., *Geometric Invariance in Computer Vision* (MIT Press, 1992).

[30] D.M. De Micheli, M. Bergamasco and P. Dario, An anthropomorphic active vision system based on a retina-like sensor, *Proc. 3rd Int. Symposium on Measurement and Control in Robotics,* Torino, Italy (September 1993).



Carlo Colombo received a degree in Electronic Engineering from the University of Florence, Italy, in January 1992. From January to September 1992 he did research on image processing and computer vision at the University of Florence. In September 1992 he joined the Advanced Robotics Technology and Systems Laboratory of the Scuola Superiore di Studi Universitari e di Perfezionamento S. Anna, Pisa, Italy, where he is currently a Ph.D. student. His research interests include the study of biological and artificial perception mechanisms for robotics applications.

Michele Rucci is a Ph.D. student at the Scuola Superiore S. Anna, Pisa, Italy. In 1991 he graduated in Electronic Engineering at the University of Florence, Italy, where he has since been engaged in research activity on medical imaging and artificial neural networks. In September 1991 he has obtained a fellowship offered by the 'Istituto per la Ricostruzione Industriale' for a two-year course in Bioengineering at the Scuola Superiore S. Anna. In the last two years he has been working in the fields of neural networks, computer vision, artificial perception and medical imaging. He has authored/coauthored more than 20 papers in journals and conference proceedings.



Paolo Dario is an Associate Professor of Biomedical Engineering at the Scuola Superiore S. Anna, Pisa, Italy. He earned a Dr. Eng. degree in Mechanical Engineering from the University of Pisa. He has been appointed as Visiting Professor by Brown University (Providence, RI), and by the Academia Sinica (Bejing, China), and was awarded a Fellowship from the Japanese Government as a Foreign Expert in Robotics. He teaches a course on Biomedical Technologies at the Scuola Superiore S. Anna and a course on Mechatronics at the School of Engineering of the University of Pisa, and is also director of the ARTS and of the MiTech labs of the Scuola Superiore S. Anna. His main research interests are in the fields of sensors and actuators for biomedical and robotic applications, and of intelligent robotic systems and microsystems. Prof. Dario is a member of the Scientific Board of the Italian National Program on Robotics and of the Italian Society of Robotics. In addition, Prof. Dario is the Chairman of the Technical Committee on Micro Robots and Cellular Robots of the IEEE Society of Robotics and Automation. Prof. Dario is a member of the Editorial Board of the *Journal of Robotic Systems*, has organized two Nato Workshops on Robotics, the 5th International Conference on Advanced Robotics (ICAR'91), and a number of Invited Sessions in international conferences. Prof. Dario is the editor of two books, and the author, or co-author, of more than 140 papers on refereed journals, books and conference proceedings, as well as the author of 8 patents (4 of which are U.S. patents).